Context-based retrieval of functional modules in protein-protein interaction networks

Dobay, Maria Pamela ; Stertz, Silke ; Delorenzi, Mauro

Abstract: Various techniques have been developed for identifying the most probable interactants of a protein under a given biological context. In this article, we dissect the effects of the choice of the protein–protein interaction network (PPI) and the manipulation of PPI settings on the network neighborhood of the influenza A virus (IAV) network, as well as hits in genome-wide small interfering RNA screen results for IAV host factors. We investigate the potential of context filtering, which uses text mining evidence linked to PPI edges, as a complement to the edge confidence scores typically provided in PPIs for filtering, for obtaining more biologically relevant network neighborhoods. Here, we estimate the maximum performance of context filtering to isolate a Kyoto Encyclopedia of Genes and Genomes (KEGG) network Ki from a union of KEGG networks and its network neighborhood. The work gives insights on the use of human PPIs in network neighborhood approaches for functional inference.

DOI: https://doi.org/10.1093/bib/bbx029

OXFORD UNIVERSITY PRESS | Briefings in Bioinformatics

Context-based retrieval of functional modules in protein-protein interaction networks

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-16-0246.R2
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Dobay, Maria Pamela; SIB Swiss Institute of Bioinformatics, Bioinformatics Core Facility Stertz, Silke; Institute of Medical Virology, University of Zurich Delorenzi, Mauro; SIB Swiss Institute of Bioinformatics, Bioinformatics Core Facility; Centre Hospitalier Universitaire Vaudois, Department of Oncology; Ludwig Center for Cancer Research of the University of Lausanne
Keywords:	Protein-protein interaction networks, Context filtering, Text mining evidence evaluation, Information content bias

Context-based retrieval of functional modules in protein-protein interaction networks

Maria Pamela Dobay[1,*], Silke Stertz[2,**], and Mauro Delorenzi[1,3,4,**]

[1]Bioinformatics Core Facility, SIB Swiss Institute of Bioinformatics Quartier Sorge, Batiment Genopode, 1015 Lausanne, Switzerland
[2]Institute of Medical Virology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich
[3]Centre Hospitalier Universitaire Vaudois, 1011 Lausanne Switzerland
[4]Ludwig Center for Cancer Research, University of Lausanne, 1066 Epalinges,
*To whom correspondence should be addressed
**Equal contribution
E-mail: Maria-Pamela.Dobay@sib.swiss

Abstract

Various techniques have been developed for identifying the most probable interactants of a protein under a given biological context. In this paper, we dissect the effects of the choice of the protein-protein interaction network (PPI) and the manipulation of PPI settings on the network neighborhood of the influenza A virus (IAV) KEGG network, as well as hits in genome-wide siRNA screen results for IAV host factors. We investigate the potential of context filtering, which uses textmining evidence linked to PPI vertices, as a complement to the edge confidence scores typically provided in PPIs for filtering, for obtaining more biologically-relevant network neighborhoods. Here, we estimate the maximum performance of context filtering to isolate a KEGG network K_i from a union of KEGG networks and its network neighborhood. The work gives insights on the use of human PPIs in network neighborhood approaches for functional inference.

Introduction

Protein function inference typically involves extensive genetic and biochemical analyses, unless good homology models exist [1, 2]. Alternatively, functions can be inferred from network associations -- viewed in the context of functional modules -- within well-characterized protein-protein interaction networks (PPIs) [3, 4]. Most of these 'protein neighborhood' inference methods that were actually used in experimentally-confirmed discovery, however, were developed using the manually-curated Mammalian Protein-Protein Interaction Database (MIPs) [5] and confirmed in Saccharomyces cerevisiae [6, 7].

Recent siRNA screens for identifying viral host factors in influenza A virus (IAV) infection have yielded numerous candidates with unknown functions. Most target prioritizations to date have been performed by finding overlapping hits across these screens [8], severely limiting the number of promising hits considered for follow-up. In the case of the genome-wide screens for IAV, the number of overlaps range from a high of 113 in at least two screens to a low of six in at least four screens; no complete overlaps are reported across screens [9]. Inferring functions for these proteins is not only important for target prioritization, but also the choice of validation assays.

We focus on the direct comparison of integrated PPIs, namely the two most recent releases of STRING [10-12] and HIPPIE [13]. STRING is a functional PPI which includes both physical interactions between proteins, as well as indirect functional interactions, such as transcriptional activation via signaling. Interactions included in STRING are inferred from multiple sources, including data from databases, experiments, textmining, genomic co-occurrence, genomic neighborhood, experimental coexpression, and gene fusion. It is benchmarked against functional groupings in the Kyoto Encyclopedia of Genes and Genomes (KEGG), which was chosen due to its manual curation, availability for multiple organisms, and coverage of different functional areas. All edges are assigned a confidence score which is the probability of finding a pair of linked proteins in a KEGG pathway, and predicted associations that are found in KEGG pathways are considered true positives [10]. In contrast, HIPPIE is a physical PPI that is restricted to experimentally-validated physical interactions between proteins, and integrates various interactions from public, curated databases as well as studies, including BioGRID, DIP, HPRD, IntAct, MINT and BIND [13]. Unlike STRING, HIPPIE explicitly removes genetic interactions, in particular those included in BioGRID. Edges in HIPPIE are associated with an interaction score, calculated as a function of the number of studies in which an interaction was detected, the number of different experimental techniques and the confidence scores linked to each of these techniques, and the number of times an interaction was found in other organisms [13].

It can thus be expected that the two PPIs would have differences in content, including in score distributions. It can likewise be expected that network neighborhoods derived from the two PPIs would be degenerate. In fact, a previous study has demonstrated that there is no general agreement between the database scores, except for 4539/31229 interactions in common in STRING and HIPPIE that were found to have high confidence scores in both PPIs [14]. This study, however, had a limited scope, mainly analyzing protein coverage, the number of interactions and network neighborhood characteristics, and does not explicitly evaluate the effects of these parameters on functional assignment or retrieval of functional modules [15]. Furthermore, the study restricted the comparison of the PPI entries to those with experimental evidence, or that were obtained from other interaction databases, which might inflate the overlaps between the databases while dramatically reducing the edges included in the study. Consequently, when using such resources for functional inference, the question remains – even for high-confidence interactions -- as to which neighbors should be prioritized for follow-up.

In this paper, we present a detailed analysis of STRING and HIPPIE in terms of their basic characteristics, including coverage, inter-PPI and inter-version concordance, and edge inclusion from primary source databases. We then checked how these differences affect the network neighborhoods retrieved for both well-characterized and less-characterized query nodes. In particular, we checked if retrieved neighbors have been implicated in the biological process of interest; in the case of STRING, we also checked the main themes of the textmining evidence associated with both the query network and its network neighborhood, and compared its overlap with the manually-curated evidence used for building the query network. This paper notably extends the scope of the

previous study by performing the comparison of PPIs on all edges, rather than a subset of edges, and checking the consequences of using different PPIs and PPI filtering methods on the retrieval of KEGG networks, as well as on the network neighborhoods of real-world examples of experimentally confirmed hits in the IAV host factor screen. Finally, we demonstrate the potential context filtering, which uses experimentally-derived or inferred annotations on PPI vertices and edges, as a complement to edge-based confidence filters for the retrieval of network neighborhoods linked to specific biological contexts.

MATERIALS AND METHODS

Protein-protein interaction (PPIs) networks We compared the physical protein-protein interaction network Human Integrated Protein-Protein Interaction rEference (HIPPIE), versions 1.7 and 1.8 [13] and the functional protein-protein interaction network STRING, versions 9.05 and 10 through its R interface (STRINGdb v.1.8.1) [10-12]. HIPPIE consolidates information from other interaction networks, as well as results from large-scale proteomics studies expected to yield information on physical interactions. HIPPIE explicitly removes genetic interactions. In contrast, STRING includes both physical protein-protein interactions from most of the databases used in HIPPIE (Supplementary Table 1), as well as functional interactions inferred from co-expression data, homology modeling, and textmining. All graphs were converted to the igraph format (R package igraph_1.0.1). Graph similarities were measured as described in Table 1. Vertices were annotated with all available gene ontology (GO) terms (biological process and cellular compartment, R packages org.Hs.eg.db_3.1.2 and GO.db_3.1.2), and when available, with the Z-score from the redundant siRNA analysis (RSA) algorithm (Z_{RSA} score) [16], a quantity reflecting the effect of gene knockdown on IAV infection [16, 17]. A lower Z_{RSA} score indicates that gene knockdown successfully inhibits a viral process of interest. For a brief description of the Z_{RSA} score calculation, please refer to the data supplement.

Edge inclusion and evaluation of PPI properties We performed general analyses that dissect graph characteristics, examine differences in the PPIs, and evaluate the effects of standard protocols that can be performed on a PPI. In particular, we checked the concordance of graph edges and topological features both globally and given a set of query nodes belonging to the same biological function; reviewed the evidence sources for establishing edge confidence scores; and evaluated the effects of confidence score filtering. We also checked if the patterns of inclusion of edges from primary databases, namely BioGrid (Releases 3.1.84, 3.2.96, 3.2.108, 3.2.120 and 3.4.131), IntAct (ftp://ftp.ebi.ac.uk/pub/databases/intact/current/all.zip, downloaded in December 2015), HPRD (Release9_062910) and MINT (2012-10-29), were disparate in STRING and HIPPIE.

Graph filtering Context filters were applied to the STRING-derived graphs, based on textmining evidence associated with its edges. Figure 1 gives an overview of all the methods used to perform context filtering on PPIs and to evaluate results.

Vertex-based context filtering We adapted the context association and filtering methods described in [15] to STRING, but with the full GO bp tree rather than GO slim.

Textmining evidence analysis and edge-based context filtering In the case of STRING, a systematic estimation of quality and scope of the textmining evidence is important. We first checked if the original references used for building a total of 35 KEGG networks from broad functional categories, including the IAV KEGG network ($KEGG_{Flu}$), 11 networks linked to $KEGG_{Flu}$, and 24 other networks from the signal transduction, cellular processes and human disease modules, overlapped with STRING textmining evidence for the same edges. All KEGG networks used were obtained as graph objects using KEGGgraph (v. 1.26), which is a wrapper for downloading KEGG pathways directly from http://www.genome.jp/kegg-bin/ in the KGML format [18]. Note that the networks extracted from this site contain updated information, and are not restricted to pathways updated in 2011 (Supplementary Figure 1).

We also checked if domain- and pathway-specific keywords are overrepresented in textmining abstracts linked to pathways examined. To isolate these context-specific terms from the abstracts, we excluded English words and stopwords, except those that are included in a biomedical corpus (https://github.com/Glutanimate/wordlist-medicalterms-en). Document term matrices, which contain the frequency of terms from the textmining abstracts, are created from filtered and stemmed text corpora (DTM, tm_0.6-2); in cases where an abstract set exceeds 10000 elements, we subsample it to a maximum size of 10000. Stemming, which reduces related words to a common root, was approximated by calculating the distance between words and merging those with a Jaro-Winkler distance greater than 0 and less than 0.1 under a common root (stringdist, v.0.9.4.1). Visualization of stemmed and merged text was performed using wordcloud_2.5. DTMs were visualized using gplots (v.2.17.0). Precision and recall rates for all edge filters explored were calculated per KEGG pathway as follows, where TP is the true positive rate; FN, the false negative rate; and FP, the false positive rate. FP_{est} consists of all extra edges in the extracted subgraphs that are not part of the original KEGG pathway. Finally, we used both expert-defined keywords and pathway-associated keywords to extract edges linked to functional groups from a high-confidence score network. Graph edges are retained if these are supported by at least one textmining evidence containing the expert-defined keywords in the abstract with a frequency that exceeds the mean for all abstracts linked to each edge.

Filtering results evaluation We use analysis of variance (ANOVA) followed by Tukey's post-hoc test to compare various pre- and post-filtering Z_{RSA} score distributions, which we use as a surrogate for evaluating how much results for IAV networks are more enriched for proteins relevant in IAV infection. Where possible, we compared the results from GO-annotation based vertex filtering or keyword-based edge filtering results to 1000 randomly-filtered subnetworks on the same number of edges as the GO- or keyword-filtered networks. We also checked pre- and post-filtered networks for comparative enrichment for textmining evidence containing expert-defined keywords (Section "Orthogonal literature evaluation'") edges. In the case of non-IAV KEGG

pathway retrieval tasks, filtering results were evaluated based on precision and recall parameters (Table 1), as well as on the enrichment of pre- and post-filtered networks for relevant textmining evidence.

Orthogonal literature evaluation As an additional metric for evaluating filtering results, we also performed independent textmining and hit identification in PubMed (rentrez, v. 1.0.0) using the combination of retained vertex names and expert-defined keywords. For edge-based filters, searches were conducted with the names of both incident vertices and expert-defined keywords. All searches were performed with the boolean AND operator. To maximize the retrieval of relevant hits (i.e. to remove matched abstracts where the keyword is found in an enumeration), we further filtered the abstracts to those where the frequency of the keywords of interest exceeded the mean for all retrieved abstracts. All orthogonal literature searches were subjected to the same processes as described in "Textmining evidence analysis and edge-based context filtering"'. Note that while the data source is the same (i.e. STRING also uses PubMed as its textmining evidence source), the method of retaining relevant edges is independent from that of STRING (thus orthogonal), as it is query-driven and is solely based on text abstracts, and not the full text.

Code availability Selected code and data files that demonstrate STRING graph manipulation, including keyword-based context filtering, can be found at https://github.com/pampernickel/flu_ppi.

RESULTS

Review of PPIs: content, edge scores and data sources

As a first step in evaluating the potential effects of PPI choice on functional inference, we first checked the degree of overlap across PPIs, and more importantly, on their most recent versions (Figure 2A-B). Based on Eq. 1, the inter-graph edge concordance between the current versions of STRING and HIPPIE is 42.5%. The inter-version concordance of the full STRING network is 33.3%, while HIPPIE inter-version concordance is 98.7%. We also checked the distribution of edge scores, which are essentially confidence estimates, for STRING and HIPPIE. The STRING confidence score is calculated as a combined probability of scores from different evidence channels, including experimental, textmining, and coexpression scores corrected for a random interaction probability, and is benchmarked against the KEGG database [10]. HIPPIE confidence scores, in contrast, reflect the reliability of the experimental evidence linked to each edge, and are calculated from the number of different studies reporting an interaction, the number of species where orthologs of the interacting proteins were found to interact experimentally, and the sum of scores from different experimental techniques used to establish an interaction [13]. While the lack of concordance between the edge score distributions of STRING and HIPPIE has been reported previously [14], and is maintained in the most recent releases (Figure 2C), the confidence scores between the versions of STRING and HIPPIE was also found change between versions

(Figure 2D).

We also checked the evolution of confidence scores for various evidence scores for each PPI. Both versions of STRING use seven primary evidence sources (Supplementary Figure 2) for the combined score calculation. Two sources, 'experiment' and 'database' could be considered scores-within-a-score; STRING experimental data are consolidated from seven interaction databases, four of which are used in HIPPIE (Supplementary Table 1). 'Database' scores, on the other hand, reflect both physical and functional interactions reported in Biocarta, BioCyc, GO, KEGG, and Reactome [10]. Among the scores, the majority of the edges in both STRING versions are supported by textmining, followed by experimental data, including gene coexpression (Supplementary Figure 2) and the majority of textmining results are associated with lower confidence scores.

Finally, we estimated the contribution of different inclusion criteria for edges from primary databases to the PPI disparity (Supplementary Table 1). At least 40% of edges from BioGrid (n_{edges}=183490) and IntAct (n_{edges} 21402) are excluded from HIPPIE and STRING (Supplementary Figure 3A). HIPPIE retains most of HPRD (n_{edges} =37039) and MINT (n_{edges} =15934). STRING, v.10, on the contrary, retains the least number of edges from these two sources, while including more from HPRD and MINT than STRING, v.9.05. Interestingly, for all databases, the peak overlap with BioGrid can be linked to the 2013 release (Supplementary Figure 3B). Note that the identities of the included edges from the primary sources are likewise different (Supplementary Figure 3C).

IAV network neighborhoods as a function of PPI choice

Given clear differences in PPIs, we next evaluated the implications of PPI choice on the neighborhood of the IAV KEGG network (Flu_{KEGG}, pathway ID: hsa05164). Figure 3A shows the topology of Flu_{KEGG} and its corresponding topologies in HIPPIE (Flu^K_H, Figure 3B) and STRING (Flu^K_S, Figure 3C). All edges incident on V Flu_{KEGG} are included. Flu^K_S contains all Flu_{KEGG}, which nonetheless comprise only 8% of E Flu^K_S. In contrast, Flu^K_H contains only 63% of Flu_{KEGG} edges, which could be expected given that HIPPIE is limited to physical interactions. Nonetheless, these edges comprise 21% of E Flu^K_H. This indicates the high incidence of extra E Flu_{KEGG}, particularly in Flu^K_S.

Information availability effects on network neighborhoods

Given the disparate sizes and scope of STRING and HIPPIE, we can expect differences in the network neighborhoods derived from these. As expected, the average network neighborhood in Flu^K_S is larger (1253 neighbors/query protein) than in Flu^K_H (124 neighbors/query protein). There is a large variability in the neighborhood sizes, depending on the query node (sd Flu^K_S = 1035, sd Flu^K_H = 160, Figure 4A).

We further examined the influence of query node identity on the network neighborhood characteristics by comparing results for the well-characterized nodes of Flu_{KEGG} and

nodes from the recently-reported IAV protein interactome, Flu_{INT}, comprised of functionally validated host proteins that interact directly with IAV proteins [17]. As Flu_{INT} is larger than Flu_{KEGG}, we obtained a random subsample with the same number of vertices as Flu_{KEGG}; Flu_{KEGG} and Flu_{INT} have no overlapping vertices. An analysis of edge and vertex characteristics show that the Flu_{KEGG} network has a significantly higher textmining score (mean Flu_{KEGG} = 328 vs. mean Flu_{INT} = 190, Figure 4B, textmining references supporting each edge (mean Flu_{KEGG} =120 vs. mean Flu_{INT} =12 references/edge, Figure 4C) and number of neighbors associated with each vertex (mean Flu_{KEGG} = 38 vs. mean Flu_{INT}=8, Figure 4D). These trends are maintained in HIPPIE (Figure 4E), albeit the neighborhood sizes per node are smaller; neighborhood sizes of each node are correlated in STRING and HIPPIE (Pearson correlation coefficient = 0.53, p-value=5.3e-07, Figure 4F). These differences roughly reflect the magnitude of information bias in PPIs -- and potentially -- the amount of information we can expect to gain from PPIs, with well-characterized queries having more neighbors than less well-characterized counterparts.

Primary reference and context concordance

We moved onto checking the relevance of the extra nodes in Flu^{K}_{S}, and to see if this information might be leveraged in the retrieval of Flu_{KEGG}. This check is possible for the STRING network, which includes textmining references. We first examined the concordance between the primary references used in generating the Flu_{KEGG}, including references associated with 11 other KEGG pathways that are linked upstream and downstream of the main Flu_{KEGG} network; and the textmining references that were associated with E (Flu_{KEGG}, Flu^{K}_{S}) as well as E Flu^{K}_{S}, E Flu_{KEGG}. Table 2 shows the subset of edge evidence sources of Flu_{KEGG} and highly related networks that are also associated as textmining evidence for Flu^{K}_{S}. Of note, only an average of 15% of the original KEGG references are matched in textmining evidence for Flu^{K}_{S}. If we expand to other KEGG networks that encompass various biological functions apart from infection, the average increases to 27.8% (Supplementary Table 2). When we checked the main content of references used to build Flu_{KEGG} (Figure 5A) and Flu^{K}_{S} (Figure 5B), we found that Flu_{KEGG} abstracts are virus-specific, while Flu^{K}_{S} abstracts are predominantly cancer- and signaling-related, with potential inclusion of some virus-linked literature.

Finally, for each vertex Flu^{K}_{H} and $Flu^{K}_{S,}$ we also checked how many have been previously linked to IAV infection (i.e. potential "true positives" for IAV involvement that were simply not included in Flu_{KEGG}) by combining the vertex name with the search terms "influenza", "virus" and "infection" in PubMed. 15% of Flu^{K}_{H} and 17% of $Flu^{K}_{S,}$ vertices were found to have at least one abstract retrieved from this combination of search terms; 46% of these vertices with evidence were found in both Flu^{K}_{H} and $Flu^{K}_{S.}$ If the search stringency is reduced by limiting the search terms to "virus" and "infection", the numbers change to 51% and 50% for Flu^{K}_{H} and Flu^{K}_{S}, respectively; of these vertices, 41% were found in both Flu^{K}_{H} and $Flu^{K}_{S.}$ In the case of Flu_{KEGG}, 64% and 88% of the vertices can be linked to IAV and general viral infection in PubMed, respectively.

Confidence score filtering of influenza networks from PPI neighborhoods

We next checked if we can retrieve Flu_{KEGG} from Flu^{K}_{S} using edge confidence score filters. Figure 6A shows the precision and recall for various confidence scores; confidence filtering on Flu^{K}_{S} has a recall exceeding 90 until a filter of 0.7 (mean recall = 91.6), but has low precision (mean=16.0), assuming a worst-case estimate that all Flu^{K}_{S} Flu_{KEGG} are false positives. Confidence score filtering does not yield a joint average recall and average precision score exceeding 50. We also checked the effects of PPI filtering on the network neighborhood of 22 IAV entry factors [19]. Unlike Flu_{KEGG} and Flu_{INT}, these factors have been linked to a very specific step in the infection process. For both STRING and HIPPIE, we retrieved the network neighborhood of each entry factor and then checked how much of the retained vertices after each step of filtering have been previously linked to endocytosis, IAV infection, or other intracellular transport processes by combining the vertex name with the search terms “endocytosis”, “influenza AND virus AND infection”, and “transport AND cytoskeleton” in PubMed. Figure 6B shows the effect of confidence score filtering on the retention of neighbors linked to terms of interest. In the case of HIPPIE, as much as 59% of the network neighborhood (V = 622) have been implicated in endocytosis, transport, or infection, with the majority (51%) specifically implicated in endocytosis. Given the distribution of HIPPIE scores (Figure 1C), the neighborhood size essentially remains constant until a score of 0.7, where the size of the neighborhood drops to 25%; note, however, that as much as 67% of this filtered, high-confidence network have been implicated in processes of interest. In contrast, a mean of 50% of a ten-fold larger network neighborhood in STRING (V = 6098) have been implicated in endocytosis, transport, or infection; except for filtering at a confidence score of 0.3, where the percentage of vertices linked to processes of interest increases to 52%, filtering does not result in an improvement of the proportion of potential true positives in the network neighborhood.

Functional module retrieval in STRING using context filters

We have shown that confidence score filters would not allow us to retrieve Flu_{KEGG} from Flu^{K}_{S}. Our results have also illustrated the inclusion of extra edges -- not necessarily false positives -- but representing non-specific or non-context-relevant relationships in retrieved networks. Various filtering techniques to restrict PPIs to those in a specific biological context were introduced in [14, 15, 20] to extract subnetworks linked to a given biological context. These filters use tissue-specific expression information [14] and (sub)cellular locations, as well as functional, disease and pathway annotations [15], and have been applied in HIPPIE, but not STRING. Here, we examine results of both vertex- and edge-trimming based protocols on the extraction of various functional modules, including KEGG networks.

Vertex annotation-based graph trimming

We examined the use of GO annotations in STRING as a first trimming protocol [14, 15]. For this scenario, we again chose to work on the neighborhood of 22 IAV entry host factors [19] rather than Flu_{KEGG} network, as this represents a more concentrated range

of functions in the IAV life cycle. The neighborhood of these 22 hits in the unfiltered STRING network is comprised of 6078 putative neighbors on 10084 edges. Filtering the neighborhood to include vertices annotated with an entry-specific GO term (Supplementary table 3, $Flu_{entry,GO}$, Figure 7A), or with a combination of GO and edge confidence score filtering ($Flu_{entry,GO,400}$, Figure 7B) result in a significant difference in the Z_{RSA} score distributions with respect to the original entry neighborhood (Figure 7C-D), indicating the enrichment for nodes that tested positive in the screen. Unlike $Flu_{entry,GO}$ and $Flu_{entry,GO,400}$, the Z_{RSA} score distribution shifts for $Flu_{entry,rand,400}$ is insignificant (Figure 7E). Finally, for each of the retained vertices, we performed a paired search in PubMed for each of the genes in $Flu_{entry,GO,400}$ using entry-associated keywords of varying specificity ('endocytosis', 'pinocytosis', 'vesicle', 'clathrin', 'trafficking', 'acidification', 'influenza', 'golgi'), and retrieved the number of abstracts that support the GO annotation. 77% of the retained vertices are associated with at least one abstract linking it to an entry- or virus-related process. Of these vertices, 61% are associated with two or more abstracts, while 8% match all the keywords (Figure 7F). In comparison, for a random, confidence-filtered network from the entry subgraph on the same number of edges ($Flu_{entry,rand,400}$), only 43% of the retained vertices are associated with at least one abstract matching inclusion criteria (data not shown).

Textmining keyword filters applied on graph edges

The most intuitive edge-based filter for textmining evidence are user-provided keywords. For this usage scenario, we used the expert-defined keywords defined in the previous section to select edges in the entry subgraph that are supported by at least one abstract containing these keywords with a frequency that exceeds the mean for all abstracts linked to each edge. This results in a graph with 734 vertices linked by 880 edges, roughly 10% of the original entry hit neighborhood ($Flu_{entry,keyword}$). As with GO-based filtering, keyword filtering results in a shift to lower Z_{RSA} scores. Corresponding term frequency profiles (Figure 8A) of abstracts linked to retained edges corroborate the enrichment of endocytosis-linked terms in this network, although a subset of the evidence is linked to neuron- and synaptic-linked processes, presumably due to abstracts that contain the keyword `vesicle'. In contrast, while edges retained in $Flu_{entry,GO}$ still bear edges mainly linked to tumor signaling, there is a reduction of apoptosis-linked literature and an increase in prominence of virus- and Ras/Rab-associated links (Figure 8B). Nonetheless, results of the orthogonal paired keyword-protein name search indicate that retained vertices in $Flu_{entry,keyword}$ and $Flu_{entry,GO,400}$ tend to be associated with literature that mentions the protein name and at least four keywords in the abstract compared to $Flu_{entry,entry,400}$ (Figure 8C). We also checked effects of a more stringent keyword-matching procedure -- in this case, requiring a combination of keywords to be matched in textmining evidence for an edge to be retained. Figure 8D indicates that combinations result in further shifts to lower Z_{RSA} values and expectedly smaller subgraphs; in the case of textmining literature linked to $Flu_{entry,STRING}$ edges, filtering constraints could only be as many as three of the keywords at a time.

Variable functional efficacy of KEGG module retrieval using context filters

To evaluate the transferability of the approach to other KEGG networks, we selected networks (n=6/26 candidate KEGG networks, Supplementary Table 1) that represent various cellular processes, and that have minimal overlapping edges (Supplementary Figure 4A). These graphs were combined and embedded within their full STRING neighborhood (7832 vertices linked by 10766 edges, Supplementary Figure 4B). We subsequently attempted to isolate the original networks as described in the methods by using a set of context-relevant keywords or gene names (Supplementary Table 4), with two filtering requirement scenarios: the first retains all edges supported by at least one reference that matches any combination of two keywords (Condition 1), while the second retention condition is more stringent by requiring at least two references that match any combination of two keywords (Condition 2).

Precision and recall calculations were made with adjusted TP, FP and FN values to reflect the number of edges that could be retrieved (Figure 9A); for instance, in the phagosome network (72 edges, 21 vertices), a total of only nine of the original edges (mPOI, Supplementary Table 5) were retrieved under condition 1. However, of the 63 unretrieved edges (uPOI), 22 did not have any textmining reference associated with the edge. Additionally, for the 41 remaining edges there were no orthogonal sources that co-mention the incident vertices with the pathway of interest, nor associated keywords. This makes the maximum TP 9, and not 72. As all 9 retrievable edges were retained, FN=0. The FP in this instance is also readjusted as a function of the total number of retrieved edges or vertices in the context-filtered network (POI_{CF}) that are neither in the original network (ePOI) nor have no evidence from an orthogonal search of involvement in the pathway of interest ($ePOI_{w/oe}$). In the phagosome network, 277/284 of ePOI do not have evidence linking the incident vertices to the phagosome directly, and are considered false positives.

Figure 9B shows that context filtering performance under both conditions 1 and 2 varies widely across pathways. The more stringent condition 2 results in an average precision gain of 1.84x and 1.7x for edges and vertices, respectively with a corresponding average recall reduction of 1.24x for edges and a negligible 1.06x for vertices. The series is too small to establish correlations between performance and pathway size or textmining evidence availability per pathway. However, the worst performance is clearly for the phagosome pathway, which has the lowest average associated textmining evidence (0.61 references/edge, as opposed to a mean of 23.7 references/edge for the other pathways considered). Textmining evidence in STRING linked to unretrieved edges in this pathway were found to be enriched in related, but non-phagosome-specific terms (e.g. ‘rab’, ‘endosome’, and less specifically, ‘cytoskeleton’, Supplementary Figure 5A). Including these in the keywords would result in better retrieval, but results would significantly overlap with the endocytosis pathway network. Retrieval of peroxisome pathway edges can likewise be improved by just by altering the keywords, specifically, by using greedy pattern matches (e.g. ‘peroxi’ instead of ‘peroxisome’, which should capture ‘peroxisome’ and ‘peroxisomal’, Supplementary Figure 5B), and for more advanced users, regular expressions (e.g. `pex(\d+)' to signify all variants of ‘pex’ followed by a number).

In the case of false positives, we again focus on the phagosome, for which the poorest performance was recorded. Only 7/284 POI_{CF} edges have orthogonal evidence linking it to both the pathway of interest and the names of the vertices incident on the edges ($ePOI_{e*}$, Supplementary Table 5). We have shown previously that increasing the stringency (i.e. condition 2) could improve precision without compromising recall too severely; these results further indicate that context filtering results might be improved by only retaining edges supported by literature that contain keywords and the incident vertices.

Finally, for false negatives that were not recovered from the graph union, but for which evidence was found in the orthogonal search, we checked if these references were published prior to the release of STRING, v.10 (i.e if the references should have been retrieved by the textmining tool). At least 53% of these references should have been retrieved for the indicated edges, assuming that the textmining run for STRING, v.10 was conducted in early 2014 (Supplementary Table 4); if late 2014 was also covered, then as much as 78% of these should have been associated with the indicated edges.

DISCUSSION

There is an extensive wealth of information contained in PPIs; however, PPI contents are linked to diverse processes. We believe that a critical step in maximizing the utility of human PPIs for novel interaction discovery, or for deducing molecular mechanisms, lies in the isolation of subnetworks linked to specific biological contexts. The most commonly available PPI filter, the confidence score, is however not designed for this purpose. Context filtering using GO and MeSH terms was recently introduced in the later versions of HIPPIE, but not in STRING, nor in the primary databases from which the PPIs were derived. The combined use of the GO filter and the STRING confidence score for extracting the IAV entry network clearly results in a subgraph that has a higher probability of context relevance than one obtained by confidence score filtering alone (Figure 7).

A potentially more intuitive approach, however, which is the use of one or more user-defined keywords to filter a graph, has never been implemented. We tested this method on STRING, taking advantage of textmining evidence associated with most of its edges. This approach maximizes the use of information already contained in the PPI. However, as its performance is dependent on any information biases in the textmining evidence, we first needed to estimate this. To our knowledge, information biases in PPIs, which include biological contexts (over)represented in a PPI and the amount of information available as a function of the query protein, have not been formally investigated before. Basic processing of textmining evidence revealed that the majority of the literature associated with STRING edges are -- perhaps expectedly -- linked to signaling and tumor biology. Filtering is required to isolate references related to other contexts. In the case of most of the KEGG networks tested, the original references are not always retrieved by STRING textmining; nonetheless, in some cases, other references

retrieved compensate for these omissions. Our results nonetheless indicate that the textmining routine of STRING could still be improved.

One of the clear challenges of this analysis is that there are no comprehensive, context-specific gold standards. In the IAV-related applications, we used changes in the distribution of Z_{RSA} as an indicator of filtering efficacy for IAV networks. The Z_{RSA} score distribution is normally distributed around a mean of -0.1, and typically retains this characteristic on random sampling.

Additionally, for network neighborhood relevance evaluation and subnetwork extraction exercises, we also used an orthogonal check, which is a combined keyword and protein name(s) search on PubMed linked by the "AND" operator, as second proxy to estimate the relevance of retained vertices and edges.

In the Flu_{KEGG} example, we see that its HIPPIE and STRING network neighborhoods (Figure 3) had a comparable proportion of vertices that can be potentially linked to viral infection in general, but these account for a maximum of ~56% of the network neighborhood. Coupled with the analysis of textmining evidence for Flu^{K}_{S} edges, one could infer that the other 40% of the neighborhood nodes were extracted from another biological context – specifically, a better-represented or more general biological context like cancer or cell signaling (Figure 5). Our results also indicate that only 40% of the nodes linked to viral infection are found in both HIPPIE and STRING neighborhoods, indicating that more useful information can be obtained by first combining the network neighborhoods from the two resources prior to context filtering.

In the IAV entry network example, keyword-based filtering allowed us to extract subgraphs supported by both STRING and orthogonal PubMed corpora enriched for entry-related terms that were not used in filtering. Extending this to other networks indicates that we can retrieve functionally-related modules that cannot otherwise be separated from each other using confidence score filtering (Supplementary Figures 4B-C). Context filter performance is nonetheless influenced by both the choice of keywords, as well as the information available in the PPI. We see these in the cases of the peroxisome and phagosome networks, respectively: for the peroxisome network, changing the keywords to a combination of patterns (e.g. 'peroxi', which captures both 'peroxisome' and 'peroxisomal') and regular expressions is expected to result in improved retrieval. In the case of the phagosome network, a third of the network cannot be retrieved due to the lack of associated textmining evidence, while other edges are linked to non-phagosome specific, but related literature. Filter performance can also improve with the use of stricter criteria that use keyword combinations instead of single keywords, or that require a minimum number of abstracts containing keyword combinations can reduce precision, without dramatically reducing recall. Our work illustrates the potential of this method, and more parameters can be systematically tested in the future to optimize keyword-based filtering in PPIs. Implementing such improvements may be eventually useful, given the downstream dependence of other prediction software such as Networkin [21], GPS 2.0 [22] and SMART [23], on information from PPIs.

CONCLUSION

Providing the option for filtering PPIs based on vertex and edge parameters, particularly keyword-based filtering with various user-manipulable stringency parameters, could be of interest in the isolation of context-specific subnetworks. This filtering option may increase the utility of PPIs in functional inference-related applications, as well as in prediction software that depend on information from PPIs.

ACKNOWLEDGEMENTS

This work was supported by SystemsX through a fellowship (2013/137) provided to MPD.

Conflict of interest statement. None declared.

BIOGRAPHICAL NOTES

Maria Pamela Dobay is a bioinformatician at the Swiss Institute of Bioinformatics (SIB). She is involved in multiple -OMICs analyses, and is particularly interested in benchmarking resources for data interpretation.

Silke Stertz is an assistant professor at the University of Zurich. Her research focuses on the interplay between influenza viruses and their host cells at the level of virus entry.

Mauro Delorenzi is the head of the Bioinformatics Core Facility at the Swiss Institute of Bioinformatics (SIB), where he performs work mainly focused on –OMICs analyses in oncology research.

KEY POINTS

- We provide a critical comparison of major protein-protein interaction networks (PPIs), STRING and HIPPIE
- We illustrate how much the choice of PPI and the previous degree of characterization of query node influences the retrieved network neighborhood size
- We show network neighborhood degeneracy and confidence score insufficiency in edge extraction tests for KEGG networks
- We implement a keyword-based context filter to extract subnetworks of interest in a given biological context

References

1. Marcotte EM, Pellegrini M, Thompson MJ et al. A combined algorithm for genome-wide prediction of protein function, Nature 1999;402:83-86.
2. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure, Nat Rev Mol Cell Biol 2007;8:995-1005.
3. Vazquez A, Flammini A, Maritan A et al. Global protein function prediction from protein-protein interaction networks, Nat Biotechnol 2003;21:697-700.
4. Peng X, Wang J, Peng W et al. Protein-protein interactions: detection, reliability assessment and applications, Brief Bioinform 2016.
5. Pagel P, Kovac S, Oesterheld M et al. The MIPS mammalian protein-protein interaction database, Bioinformatics 2005;21:832-834.
6. Hishigaki H, Nakai K, Ono T et al. Assessment of prediction accuracy of protein function from protein--protein interaction data, Yeast 2001;18:523-531.
7. Trivodaliev K, Bogojeska A, Kocarev L. Exploring function prediction in protein interaction networks via clustering methods, PLoS One 2014;9:e99755.
8. Pohl MO, Edinger TO, Stertz S. Prolidase is required for early trafficking events during influenza A virus entry, J Virol 2014;88:11271-11283.
9. Chou YC, Lai MM, Wu YC et al. Variations in genome-wide RNAi screens: lessons from influenza research, J Clin Bioinforma 2015;5:2.
10. von Mering C, Jensen LJ, Snel B et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms, Nucleic Acids Res 2005;33:D433-437.
11. Franceschini A, Szklarczyk D, Frankild S et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Res 2013;41:D808-815.
12. Szklarczyk D, Franceschini A, Wyder S et al. STRING v10: protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res 2015;43:D447-452.
13. Schaefer MH, Fontaine JF, Vinayagam A et al. HIPPIE: Integrating protein interaction networks with experiment based quality scores, PLoS One 2012;7:e31826.
14. Lopes TJ, Schaefer M, Shoemaker J et al. Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases, Bioinformatics 2011;27:2414-2421.
15. Schaefer MH, Lopes TJ, Mah N et al. Adding protein context to the human protein-protein interaction network to reveal meaningful interactions, PLoS Comput Biol 2013;9:e1002860.
16. Konig R, Chiang CY, Tu BP et al. A probability-based approach for the analysis of large-scale RNAi screens, Nat Methods 2007;4:847-849.
17. Tripathi S, Pohl MO, Zhou Y et al. Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding, Cell Host Microbe 2015;18:723-735.
18. Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor, Bioinformatics 2009;25:1470-1471.
19. Edinger TO, Pohl MO, Stertz S. Entry of influenza A virus: host factors and antiviral targets, J Gen Virol 2014;95:263-277.

20. Lee SA, Chan CH, Chen TC et al. POINeT: protein interactome with sub-network analysis and hub prioritization, BMC Bioinformatics 2009;10:114.
21. Linding R, Jensen LJ, Pasculescu A et al. NetworKIN: a resource for exploring cellular phosphorylation networks, Nucleic Acids Res 2008;36:D695-699.
22. Xue Y, Ren J, Gao X et al. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy, Mol Cell Proteomics 2008;7:1598-1608.
23. Schultz J, Milpetz F, Bork P et al. SMART, a simple modular architecture research tool: identification of signaling domains, Proc Natl Acad Sci U S A 1998;95:5857-5864.

FIGURE CAPTIONS

Figure 1. Overview of graph filtering and filtering evaluation methods. The typical workflow begins with the retrieval of the network neighborhood of a query node in a PPI. The neighborhood is filtered for the most relevant hits based on vertex or edge attributes. Note that the solution for these filters can be degenerate. Filtering results are compared against unfiltered or randomly filtered counterparts, and are evaluated using orthogonal sources, including shifts in distributions of experimental scores, in this case, the Z_{RSA} scores, as well as reference counts from independent searches that indicate the potential involve- ment of the retained vertices in the process of interest.

Figure 2. Venn diagram of common edges (A) and vertices (B) across the PPIs considered indicate that only a very small proportion of features are common between STRING and HIPPIE. Edge confidence score distributions of HIPPIE and both versions of STRING show minimal concordance (C). Density distributions of all edges and edges found in both HIPPIE and STRING are shown as a function of the confidence scores of each graph. Note that edges found in both HIPPIE and STRING tend to be associated with higher STRING confidence scores. Edge evidence scores for the current and immediate previous versions of both PPIs indicate marked instances of edge confidence score changes (D).

Figure 3. Network topology of the manually-curated influenza A KEGG network (Flu_{KEGG}, A) and corresponding topologies in HIPPIE (B) and STRING (C). Node sizes are inversely proportional to the Z_{RSA} scores; more critical host factors are represented as larger nodes. Confidence scores in HIPPIE and STRING are indicated.

Figure 4: Network neighborhood of the influenza A KEGG network (Flu_{KEGG}) based on unfiltered STRING and HIPPIE shows network neighborhood variance as a function of the query node (A). STRING network neighborhoods (B-D) of Flu_{KEGG} compared to the host protein interactome (Flu_{INT}) show significantly higher textmining scores (B) and textmining references per edge (C) in STRING. The trend towards a higher number of neighbors linked to better-characterized queries appear consistently in STRING (D) and HIPPIE (E), albeit HIPPIE has smaller neighborhood sizes. There is a significant correlation (Pearson correlation coefficient = 0.53, p.val = $5.28e^{-07}$) between the number of neighbors per node in HIPPIE and STRING (F).

Figure 5: Most frequent context-specific words from the abstracts of Flu_{KEGG} references (A), of corresponding textmining sources linked to $Flu_{KEGG,STRING}$ (B) and for all links of Flu^{K}_{S}.

Figure 6. Confidence score filtering in STRING and HIPPIE. Recall and precision in the retrieval of the original $KEGG_{FLU}$ in STRING as a function of confidence score filtering

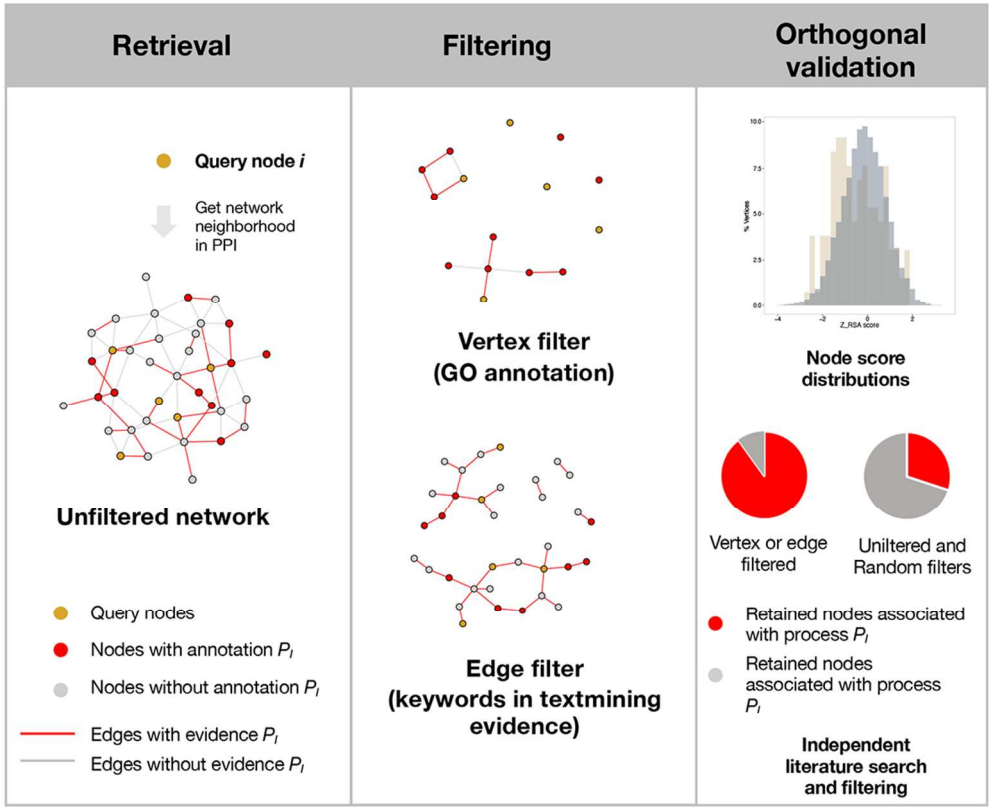
(A). Effects of confidence score filtering on the network neighborhood of entry factors in HIPPIE and STRING (B). As much as 59% of the HIPPIE network neighborhood (V = 622) have been implicated in endocytosis (up to 51%), transport, or infection. The neighborhood essentially remains static until a score of 0.7, where as much as 67% of the high-confidence network have been implicated in processes of interest. However, the size of the neighborhood drops to 25%. In contrast, only 50% of vertices of a ten-fold larger network neighborhood in STRING (V = 6098) have been implicated in endocytosis, transport, or infection; except for minor improvements at a confidence score of 0.3, filtering does not result in an improvement of the proportion of potential true positives in the network neighborhood.

Figure 7. GO annotation-filtered ($Flu_{entry,GO}$, A) and GO annotation- and confidence-filtered ($Flu_{entry,GO400}$, B) neighborhood graphs for 22 IAV entry factors. Z_{RSA} score distributions for both $Flu_{entry,GO}$ (C, $median_{ZRSA}$ = -0.27; Tukey's post-hoc test adj.p.val = 0.05) and $Flu_{entry,GO400}$ (D, $median_{ZRSA}$ = -0.48; Tukey's post-hoc test adj.p.val. = $1.1e^{-03}$) in (A) and (B) show a shift to lower Z_{RSA} scores with respect to the original entry network neighborhood (median $_{ZRSA}$ = -0.14), which indicates an enrichment for putative host factors. A representative, randomly-filtered subgraph of the entry network on the same number of nodes as $Flu_{entry,GO400}$ does not result in a similar shift in the Z_{RSA} score distribution (E, $median_{ZRSA}$ = -0.08, Tukey's post-hoc test adj.p.val. = 0.93). Orthogonal search on Pubmed using keyword-protein name pairs indicates that 77% of the retained vertices are supported by at least one abstract containing the keyword-protein name; of these, 8% have a match for all keywords (F).

Figure 8. Textmining evidence frequency profiles linked to retained edges in keyword (A) and GO-filtered (B) graphs show a shift from tumor signaling- and apoptosis-linked literature, and an increase in endocytosis- related terms, including those not used in filtering (e.g. 'Rab', 'dynamin'). While filtering results are not exactly the same, both filters result in a significantly higher retention of vertices that associated with literature from an orthogonal source that contain multiple entry-related keywords than confidence-filtered networks (C). More stringent edge-based keyword filtering, which requires multiple keyword matches, results in expectedly smaller graphs with lower mean Z_{RSA} scores than the original entry subgraph (dotted line, D).

Figure 9. Performance of context filtering under two conditions for selected KEGG pathways. Method detailing the adjustment of true positive (TP), false positive (FP) and false negative (FN) rates for recall and precision calculations (A). Matches (mPOI) between the pathway of interest (POI) and the retrieved parts of the POI after context filtering (POI_{CF}), together with extra edges in POICF that were found to have supporting, orthogonal evidence ($ePOI_e$) are considered as TP in the calculations. Extra edges without supporting orthogonal evidence ($ePOI_{w/oe}$) are considered FP, while unretrieved edges with supporting evidence (orthogonal or linked to filtered edges, $uPOI_e$) are false negatives. All unretrieved edges without any associated evidence ($uPOI_{w/oe}$) are not

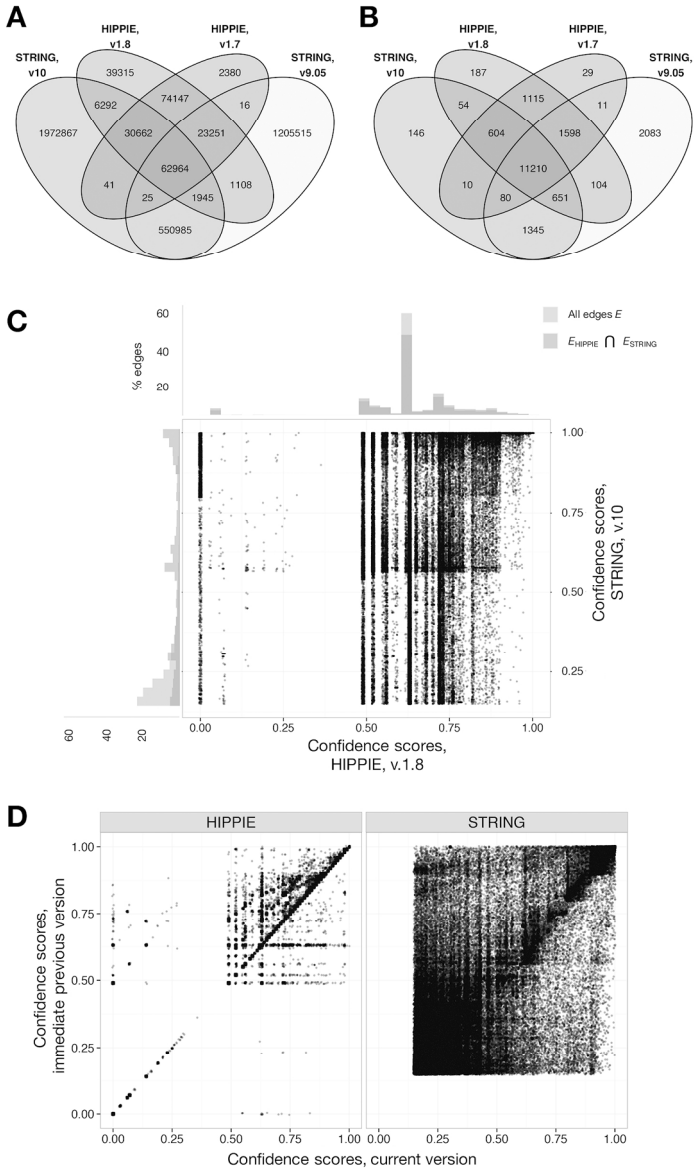
considered in precision and recall calculations. Precision and recall calculations indicate the variability of keyword-filtering performance across pathways (B). Note that stricter criteria, which requires at least two references with context-relevant evidence to support retained edges generally improves precision without a generally massive tradeoff in recall.



Overview of graph filtering and filtering evaluation methods. The typical workflow begins with the retrieval of the network neighborhood of a query node in a PPI. The neighborhood is filtered for the most relevant hits based on vertex or edge attributes. Note that the solution for these filters can be degenerate. Filtering results are compared against unfiltered or randomly filtered counterparts, and are evaluated using orthogonal sources, including shifts in distributions of experimental scores, in this case, the ZRSA scores, as well as reference counts from independent searches that indicate the potential involve- ment of the retained vertices in the process of interest.

Figure 1

151x124mm (200 x 200 DPI)



Venn diagram of common edges (A) and vertices (B) across the PPIs considered indicate that only a very small proportion of features are common between STRING and HIPPIE. Edge confidence score distributions of HIPPIE and both versions of STRING show minimal concordance (C). Density distributions of all edges and edges found in both HIPPIE and STRING are shown as a function of the confidence scores of each graph. Note that edges found in both HIPPIE and STRING tend to be associated with higher STRING confidence scores. Edge evidence scores for the current and immediate previous versions of both PPIs indicate marked instances of edge confidence score changes (D).

Figure 2

381x635mm (100 x 100 DPI)

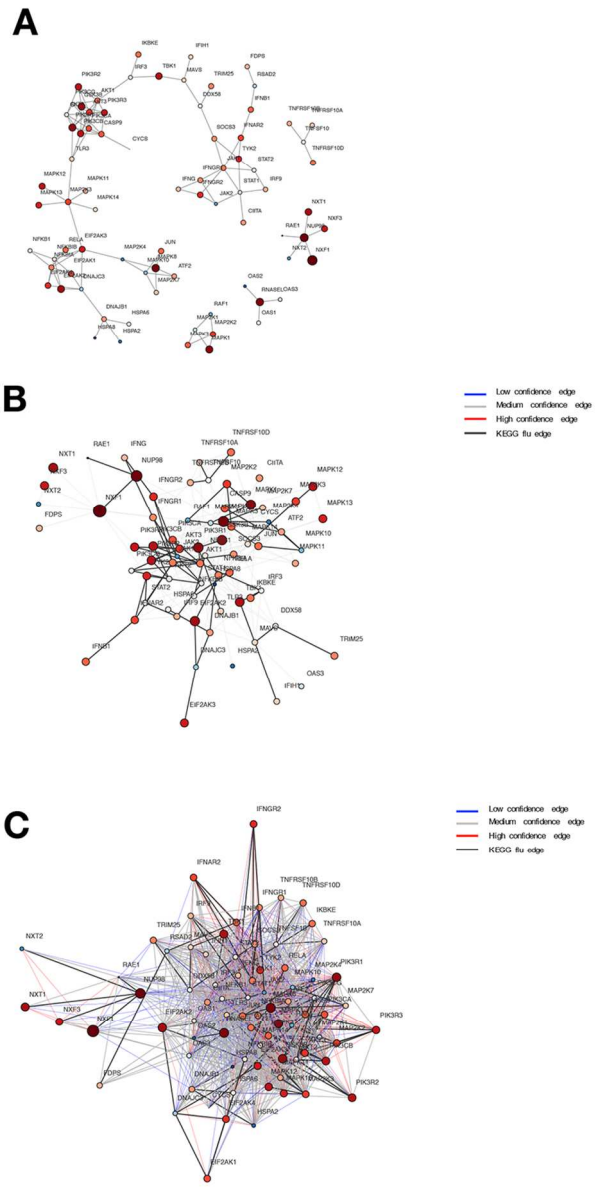
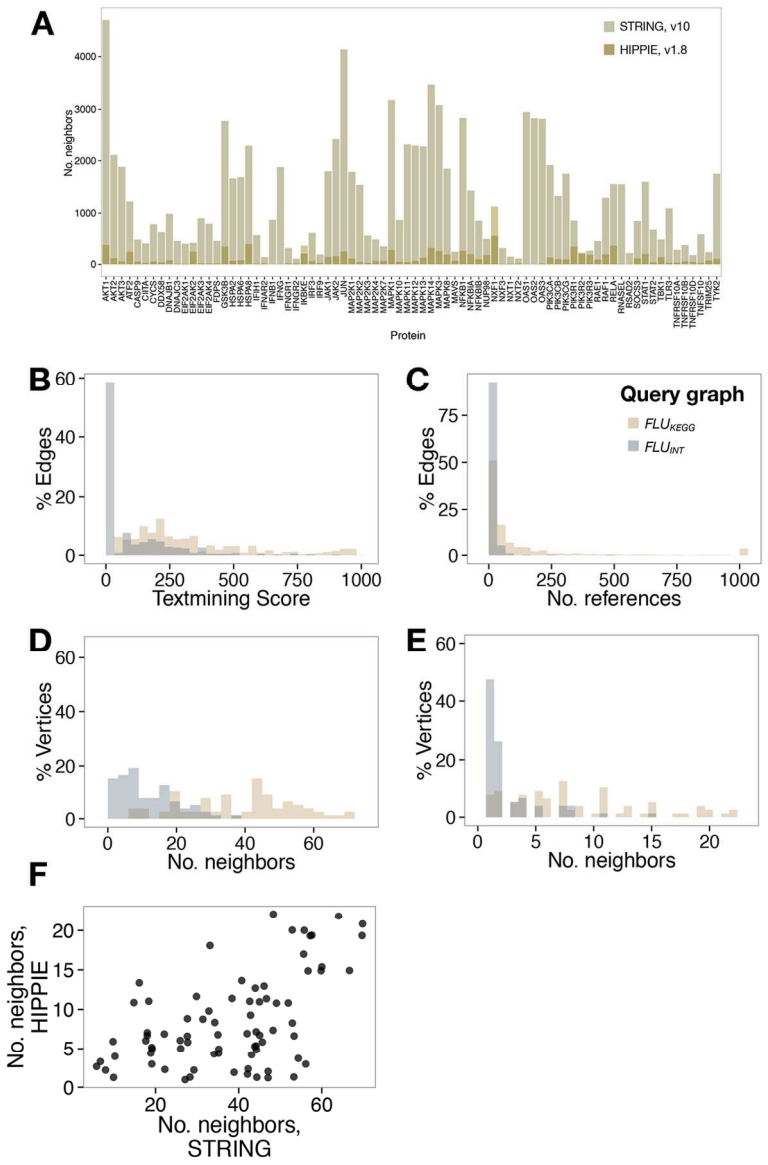


Figure 3. Network topology of the manually-curated influenza A KEGG net- work (FluKEGG, A) and corresponding topologies in HIPPIE (B) and STRING (C). Node sizes are inversely proportional to the ZRSA scores; more critical host factors are represented as larger nodes. Confidence scores in HIPPIE and STRING are indicated.

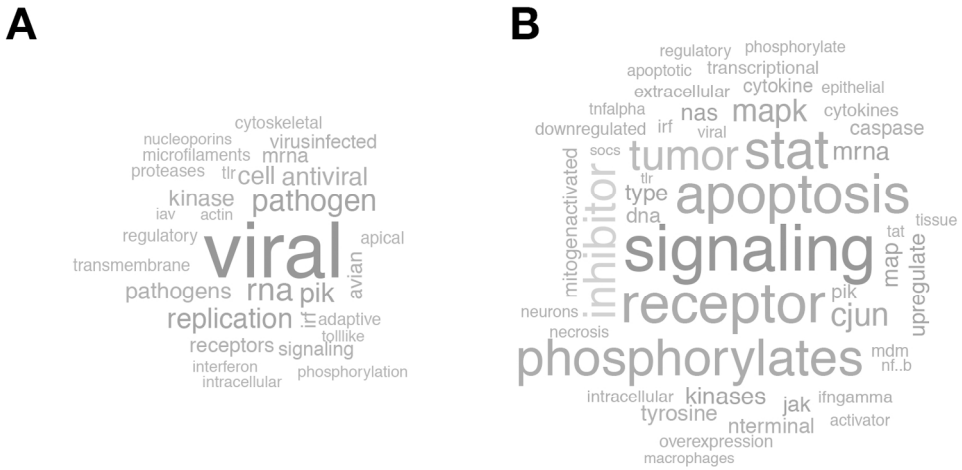
Figure 3
106x209mm (200 x 200 DPI)



Network neighborhood of the influenza A KEGG network (FluKEGG) based on unfiltered STRING and HIPPIE shows network neighborhood variance as a function of the query node (A). STRING network neighborhoods (B-D) of FluKEGG compared to the host protein interactome (FluINT) show significantly higher textmining scores (B) and textmining references per edge (C) in STRING. The trend towards a higher number of neighbors linked to better-characterized queries appear consistently in STRING (D) and HIPPIE (E), albeit HIPPIE has smaller neighborhood sizes. There is a significant cor- relation (Pearson correlation coefficient = 0.53, p.val = 5.28e-07) between the number of neighbors per node in HIPPIE and STRING (F).

Figure 4

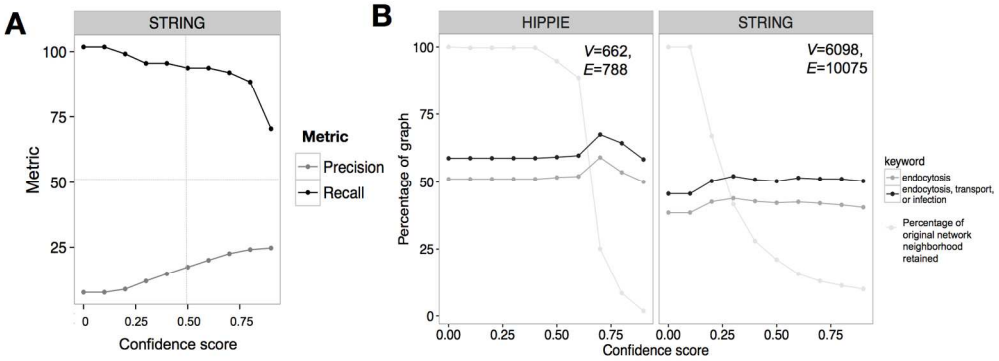
157x242mm (200 x 200 DPI)



Most frequent context-specific words from the abstracts of FluKEGG references (A), of corresponding textmining sources linked to FluKEGG,STRING (B) and for all links of FluKS.

Figure 5

150x76mm (300 x 300 DPI)



Confidence score filtering in STRING and HIPPIE. Recall and precision in the retrieval of the original KEGGFLU in STRING as a function of confidence score filtering (A). Effects of confidence score filtering on the network neighborhood of entry factors in HIPPIE and STRING (B). As much as 59% of the HIPPIE network neighborhood (V = 622) have been implicated in endocytosis (up to 51%), transport, or infection. The neighborhood essentially remains static until a score of 0.7, where as much as 67% of the high-confidence network have been implicated in processes of interest. However, the size of the neighborhood drops to 25%. In contrast, only 50% of vertices of a ten-fold larger network neighborhood in STRING (V = 6098) have been implicated in endocytosis, transport, or infection; except for minor improvements at a confidence score of 0.3, filtering does not result in an improvement of the proportion of potential true positives in the network neighborhood.

Figure 6
324x119mm (150 x 150 DPI)

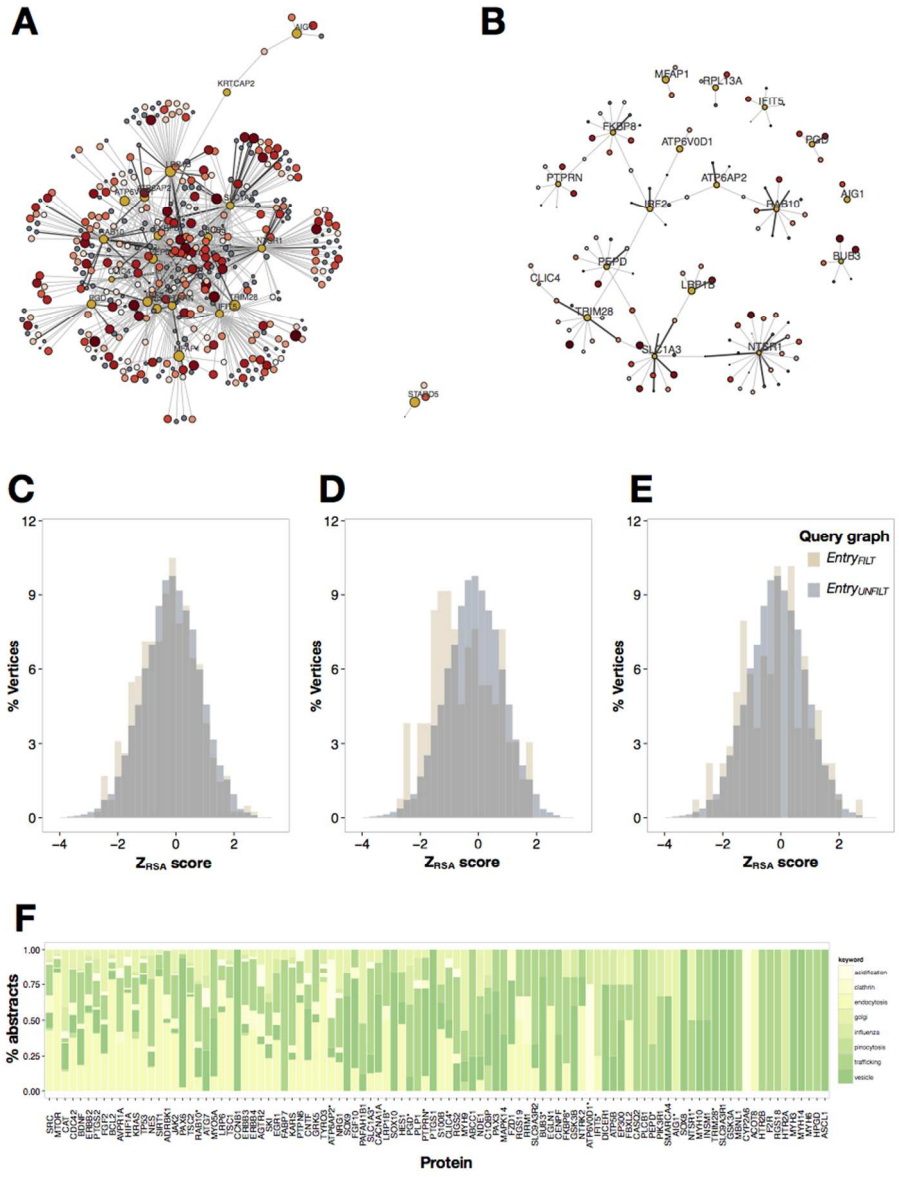
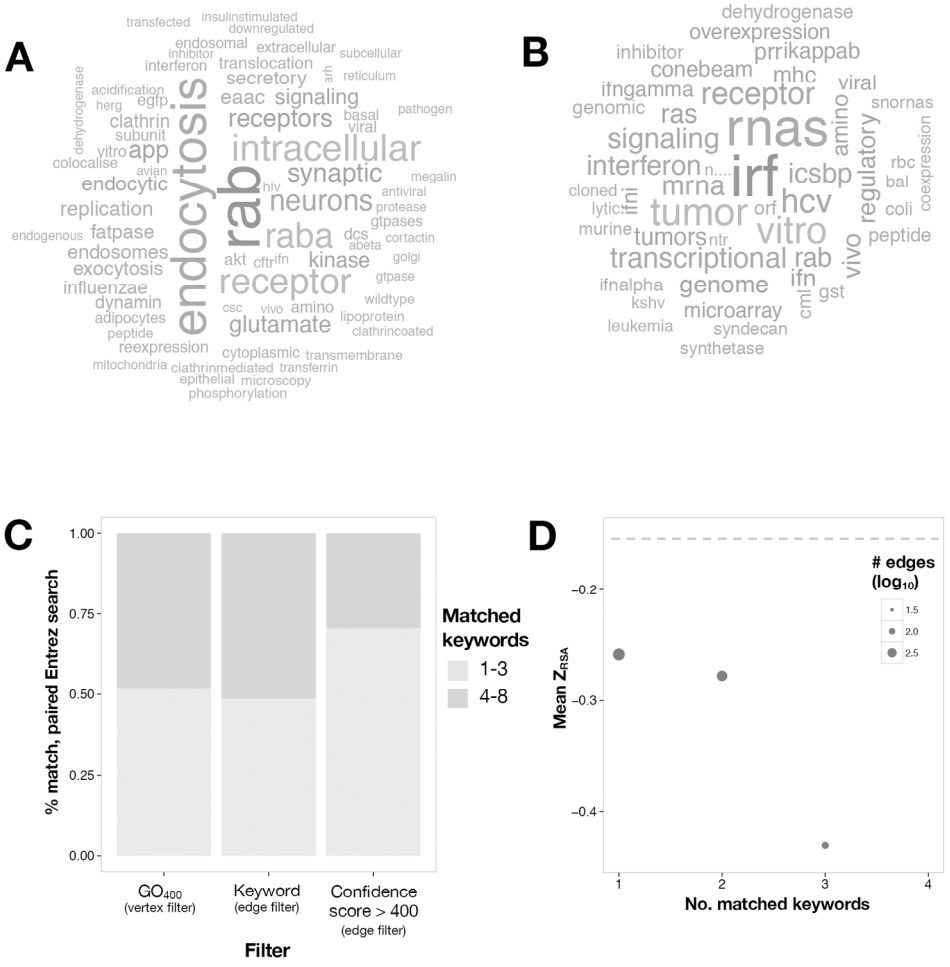


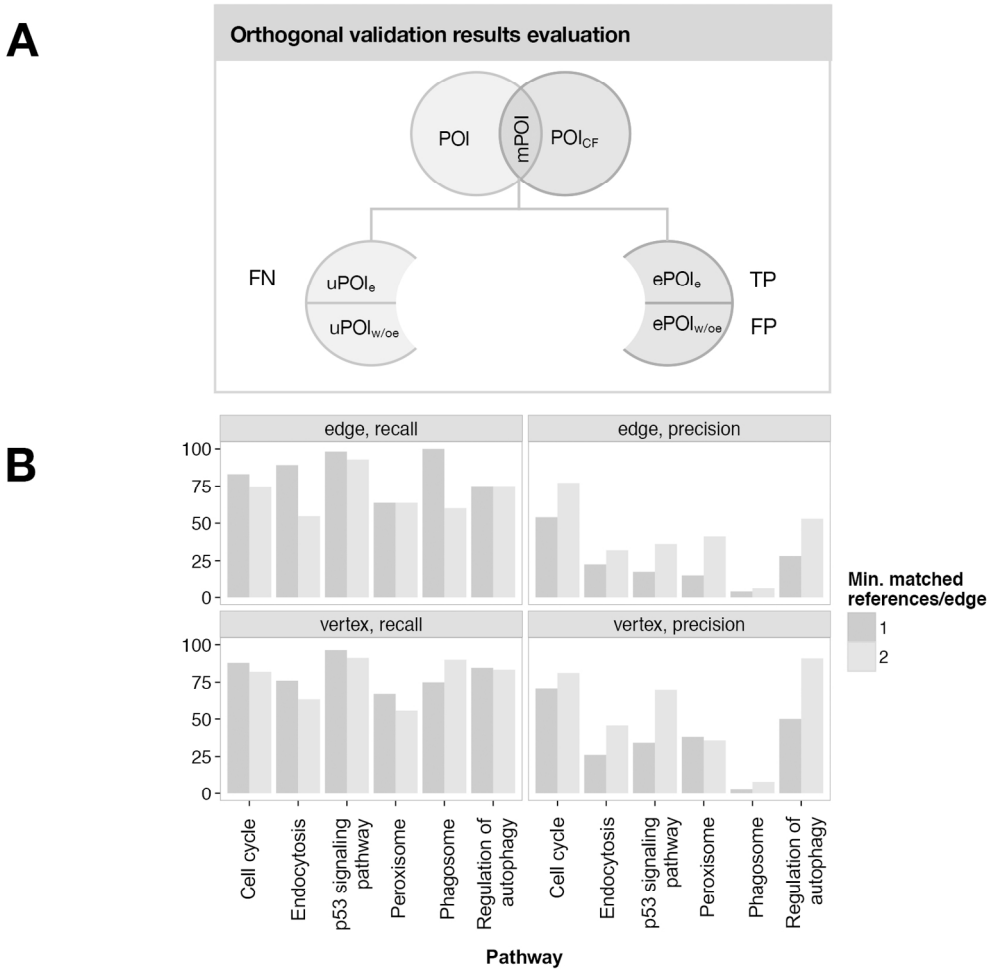
Figure 7. GO annotation-filtered (Fluentry,GO, A) and GO annotation- and confidence-filtered (Fluentry,GO400, B) neighborhood graphs for 22 IAV entry factors. ZRSA score distributions for both Fluentry,GO (C, medianZRSA = -0.27; Tukey's post-hoc test adj.p.val = 0.05) and Fluentry,GO400 (D, medianZRSA = -0.48; Tukey's post-hoc test adj.p.val. = 1.1e-03) in (A) and (B) show a shift to lower ZRSA scores with respect to the original entry network neighborhood (median ZRSA = -0.14), which indicates an enrichment for putative host factors. A representative, randomly-filtered subgraph of the entry network on the same number of nodes as Fluentry,GO400 does not result in a similar shift in the ZRSA score distribution (E, medianZRSA = -0.08, Tukey's post-hoc test adj.p.val. = 0.93). Orthogonal search on Pubmed using keyword-protein name pairs indicates that 77% of the retained vertices are supported by at least one abstract containing the keyword-protein name; of these, 8% have a match for all keywords (F).
Figure 7



Textmining evidence frequency profiles linked to retained edges in keyword (A) and GO-filtered (B) graphs show a shift from tumor signaling- and apoptosis-linked literature, and an increase in endocytosis- related terms, including those not used in filtering (e.g. 'Rab', 'dynamin'). While filtering results are not exactly the same, both filters result in a significantly higher retention of vertices that associated with literature from an orthogonal source that contain multiple entry-related keywords than confidence-filtered networks (C). More stringent edge-based keyword filtering, which requires multiple keyword matches, results in expectedly smaller graphs with lower mean ZRSA scores than the original entry subgraph (dotted line, D).

Figure 8

162x164mm (300 x 300 DPI)



Performance of context filtering under two conditions for selected KEGG pathways. Method detailing the adjustment of true positive (TP), false positive (FP) and false negative (FN) rates for recall and precision calculations (A). Matches (mPOI) between the pathway of interest (POI) and the retrieved parts of the POI after context filtering (POICF), together with extra edges in POICF that were found to have supporting, orthogonal evidence (ePOIe) are considered as TP in the calculations. Extra edges without supporting orthogonal evidence (ePOIw/oe) are considered FP, while unretrieved edges with supporting evidence (orthogonal or linked to filtered edges, uPOIe) are false negatives. All unretrieved edges without any associated evidence (uPOIw/oe) are not considered in precision and recall calculations. Precision and recall calculations indicate the variability of keyword-filtering performance across pathways (B). Note that stricter criteria, which requires at least two references with context-relevant evidence to support retained edges generally improves precision without a generally massive tradeoff in recall.

Figure 9
143x145mm (300 x 300 DPI)

Table 1. Formulae for comparing graphs and assessing effects of graph operations

Description	Formula
Graph similarity (Eq. 1)	$\frac{G1_{edges} \bigcap G2_{edges}}{min(G1_{edges}, G2_{edges})}$
Recall (Eq. 2)	$\frac{TP}{(TP + FN)}$
Precision (Eq. 3)	$\frac{TP}{(TP + FP_{est})}$

Table 2. Primary reference concordance between KEGGs and corresponding textmining edge evidence in the KEGG IAV network in STRING

KEGG ID	Pathway	KEGG refs.	Matched in STRING
hsa05164	Influenza A	20	4
hsa04621	Nod-like	16	1
hsa04620	Toll-like	12	3
hsa04622	Rig-1-like	12	9
hsa04630	JAK-STAT	11	2
hsa04010	MAPK	9	1
hsa04144	Endocytosis	10	0
hsa05416	Viral myocarditis	14	0
hsa03013	RNA transport	9	0
hsa03015	mRNA surveillance	5	0
hsa04210	Apoptosis	32	4

Supplementary Information

Dobay/Stertz/Delorenzi

December 6, 2016

1 Glossary of terms

Corpus A structured set of texts that are used in statistical analyses of document content.

Document Term Matrix (DTM) A matrix that describes the frequency of occurrence of terms or word stems (see **'Stem (also stemming, stemmed)'**) in a collection of documents.

Jaro-Winkler distance A distance d characterizing the difference between two character strings, *s1* and *s2*, given by:

$$d_w = d_j + lp * (1 - d_j) \tag{1}$$

where

$$\mathrm{d}_j = \begin{cases} 0 & m = 0 \\ \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}) & m! = 0 \end{cases}$$

Here, m is the number of matching characters; two characters from s_1 and s_2 are considered matching if they are the not farther than:

$$\left\lfloor \frac{max(|s_1|, |s_2|)}{2} \right\rfloor - 1 \tag{2}$$

t is 0.5x the number of transpositions, defined by the number of matching characters in a different sequence order (e.g. CAR and RACE have m=3, as 'C', 'A' and 'R' are matched in both strings with an order distance not exceeding 1; consequently, $\mathrm{t}_{CAR,RACE}$=0.5*3=1.5), l is the length of the common characters from the start of the string up to a maximum of four characters; and p is a scaling factor, $0 < p < 0.25$, for adjusting the score as a function of the common prefix length.

Stem (also stemming, stemmed) Words in base or root form, e.g. for the words 'stem' is the stem for 'stemming', 'stemmed' and 'stems'.

Stopwords Words that do not contain specific/significant information for use in search queries.

Z_{RSA}. A score reflecting the effect of gene knockdown on IAV infection[1, 2]. A lower Z_{RSA} score indicates that gene knockdown successfully inhibits a viral process of interest. Briefly, all siRNAs are ranked based on their knockdown potency in descending order, with the premise that true positive hits would have multiple siRNAs positioned at the top. Potency is defined by setting two arbitrary activity thresholds, A_{min} and A_{max}, which are thresholds for defining which wells show activity. Active wells are defined as wells with an activity exceeding A_{min} are considered active, while wells less active than A_{max} are considered inactive. The probability of calling activity for wells for each gene g randomly is calculated from a random selection of r_a wells as:

$$P_a = P(N, n, r_a, r = a) \tag{3}$$

Where P is the cumulated hypergeometric function distribution. $P(N, n, m, r)$ was calculated for all genes. For details of the calculation, refer to [1].

2 Tables

Table 1: Primary database sources for selected PPIs. Five out of nine of the primary database sources were used in both STRING and HIPPIE.

Source	STRING	HIPPIE
BIND	y	y
BioGRID	y	y
DIP	y	n
HPRD	y	y
I2D	n	y
IntAct	y	y
MINT	y	y
MIPS	n	y
PID	y	n

KEGG ID	Pathway	KEGG refs.	Matched in STRING
hsa05200	Pathways in cancer	159	45
hsa05223	Non-small cell lung cancer	22	9
hsa05202	Transcriptional misregulation in cancer	68	25
hsa05210	Colorectal cancer	20	5
hsa05211	Renal cell carcinoma	8	2
hsa05212	Pancreatic cancer	14	3
hsa05214	Glioma	9	1
hsa05216	Thyroid cancer	8	3
hsa05218	Melanoma	13	4
hsa05215	Prostate cancer	11	3
hsa05221	Acute myeloid leukemia	10	2
hsa05166	HTLV-I infection	20	6
hsa05164	Influenza A	20	3
hsa05168	Herpes simplex infection	30	9
hsa05160	Hepatitis C	14	5
hsa05161	Hepatitis B	37	20
hsa05162	Measles	19	7
hsa04115	p53 signaling pathway	10	6
hsa04310	Wnt signaling pathway	10	2
hsa04012	ErbB signaling pathway	15	8
hsa04144	Endocytosis	10	0
hsa04145	Phagosome	7	1
hsa04146	Peroxisome	4	0
hsa04140	Regulation of autophagy	9	0
hsa04110	Cell cycle	22	13
hsa04210	Apoptosis	32	5

Table 2: Primary reference concordance between KEGGs and corresponding textmining edge evidence in STRING. 'KEGG refs.' refer to the number of KEGG references used to create a given KEGG network; 'Matched in STRING' refer to the subset of KEGG refs. that were associated as textmining evidence with the relevant KEGG edges in STRING, v.10.

3 Figures

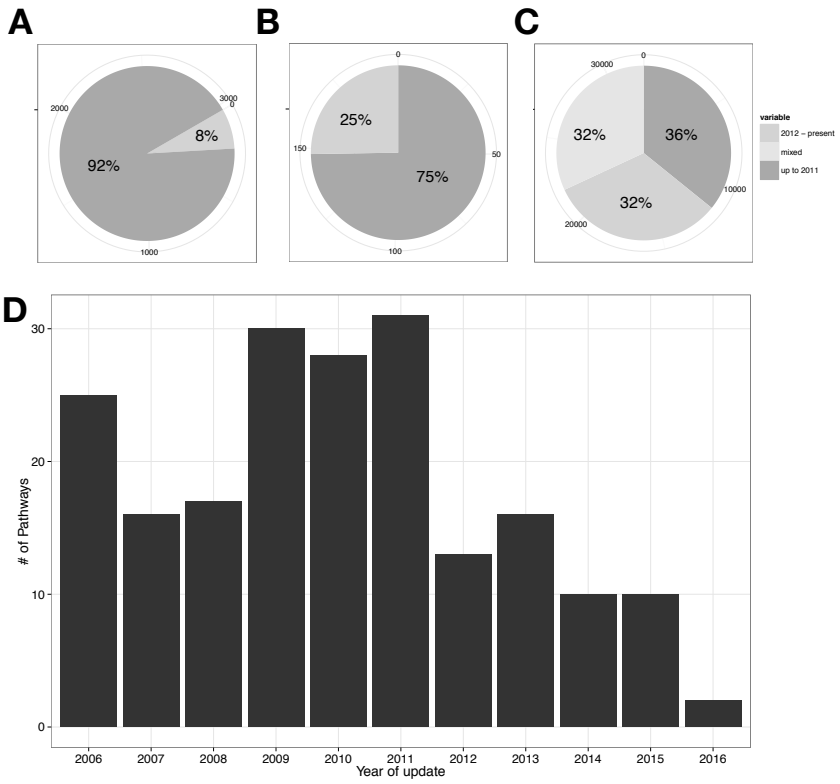


Figure 1: KEGG version updates. Of the primary references used to build KEGG networks, 8% are from 2012-2016 (A), with 25% of all pathways having at least one reference from 2012 to the present (B). Only 36% of the edges are linked to pathways exclusively supported by references from 2011 (C). Finally, if we consider the KEGG pathway map update history, we again see that a total of 25% of the pathways have been updated from 2012 to 2016, while the rest of the pathways have remained static (i.e, last update in 2011, D).

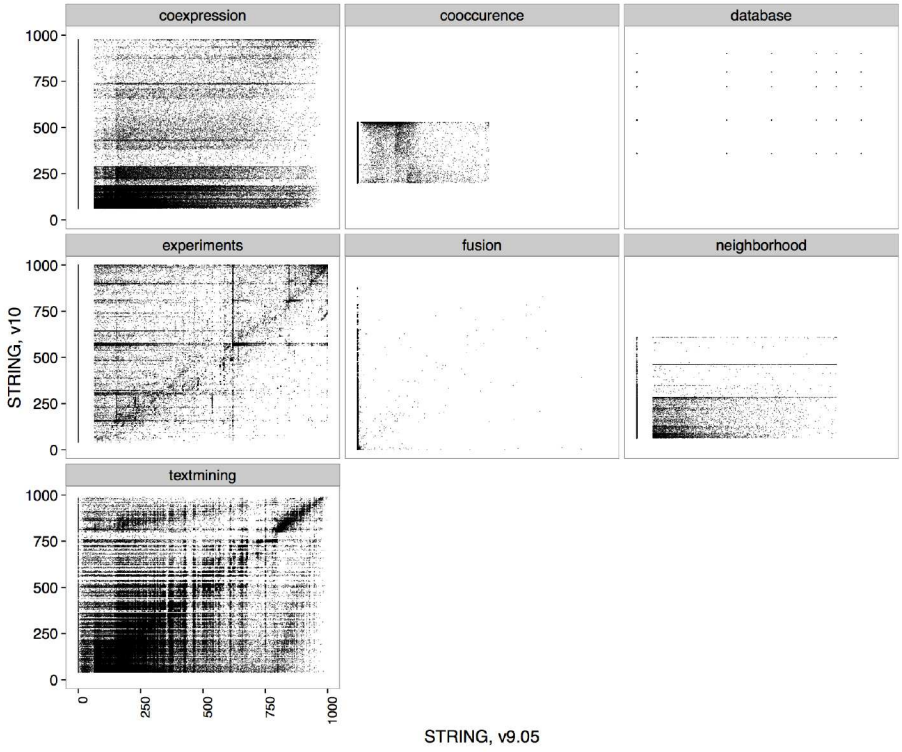


Figure 2: Inter-version concordance of STRING v.9.05 and v.10 evidence scores. Scores for individual sources of evidence used in calculating the STRING edge confidence score have changed between v.9.05 and v.10; these changes are not limited to low- or moderate-scoring edge (i.e. score < 400), but also involve downgrades of high-confidence scores. Generalizable patterns are reduction of coexpression scores in v.10, and an increase in fusion scores. Most textmining and neighborhood evidence changes are restricted to low- and moderate-scoring edges.

Pathway	Keywords
Endocytosis	endocytosis, uptake, entry, transport, endosome, clathrin, rab5, caveolin, rab7
Phagosome	phagocytosis, phagosome, acid, lysosome, endoplasmic reticulum, phagocyt*, tubulin
Peroxisome	peroxisome, fatty acid, catabolism, peroxide, pex14
Regulation of autophagy	autophag*, atg, misfolded, lysosome
p53 signaling pathway	p53, apoptosis, caspase
Cell cycle	cell cycle, cyclin, mitosis, cytokinesis, mdm, cdk

Table 3: Keywords used for retrieving KEGG subnetworks by context filtering. These include keyword roots (e.g. phagocyt*) that could be used for greedy pattern matching.

PMID	Relevant KEGG edge	Pub. date	Expected in STRING v.10 (y/n)
22082872	ATG5-ATG16L1	Dec-11	y
24899049	ATG3-ATG7	Aug-14	y
21193819	ATG3-ATG7	Dec-10	y
22576012	CDKN1A-TP53	Jul-12	y
25955014	PIK3C3-PIK3R4	May-15	n
25568150	PIK3C3-PIK3R4	Jan-15	n
24879154	PIK3C3-PIK3R4	Jun-14	y
26649827	ATG5-ATG16L1	Dec-15	n
25787994	ATG5-ATG16L1	Mar-15	n
25578879	ATG5-ATG16L1	Jan-15	n
25495476	ATG5-ATG16L1	Dec-14	y
25484075	ATG5-ATG16L1	Dec-14	y
25484072	ATG5-ATG16L1	Jan-15	n
25046113	ATG5-ATG16L1	Sep-14	y
24954904	ATG5-ATG16L1	Jul-14	y
24086718	ATG5-ATG16L1	Sep-13	y
22874553	ATG5-ATG16L1	Nov-12	y
20639694	ATG5-ATG16L1	Aug-10	y
18670194	ATG5-ATG16L1	Aug-08	y
18398292	ATG5-ATG16L1	May-08	y
23112293	GABARAPL2-ATG7; ATG3-ATG7	Nov-12	y
24474777	C3-TLR4	Jan-14	y
10559466	ITGAX-ITGAL	Nov-99	y
26043790	MAP1LC3B-MTOR	Jul-15	n
25951193	MAP1LC3B-MTOR	Jan-15	n
24991833	MAP1LC3B-MTOR	Aug-14	y
24598403	MAP1LC3B-MTOR	May-14	y
23585825	MAP1LC3B-MTOR; SQSTM1 -NFE2L2	Apr-13	y
22679478	MAP1LC3B-MTOR	May-12	y
24036548	ATG3-ATG7	Oct-13	y
23388496	ATG3-ATG7	Apr-13	y
22325599	ATG3-ATG7	Feb-12	y
22024753	ATG3-ATG7	Dec-11	y
20723759	ATG3-ATG7	Aug-10	y
16300744	ATG3-ATG7	Jan-06	y
26729618	BAX3-CASP3	Jan-16	n
25319231	BAX3-CASP3	Dec-14	y
25127907	BAX3-CASP3	Dec-14	y
24427275	BAX3-CASP3	Jan-14	y
21695150	BAX3-CASP3	Jun-11	y
18025862	BAX3-CASP3	Jan-08	y
23440701	GABARAP-ATG7	Jun-13	y
26046590	SQSTM1-NFE2L2	Jul-15	n
25049227	SQSTM1-NFE2L2	Sep-14	y
23989536	SQSTM1-NFE2L2	Oct-13	y

Table 4: References linked to unretrieved edges in the phagosome network obtained from an orthogonal search of PubMed. All references marked 'y' were published before the v.10 release of STRING (released Apr 12, 2015), and are examples of PubMed abstracts that should have been included as textmining evidence for the indicated edge. This indicates that the retrieval rate for the KEGG phagosome network in STRING could be improved by using a better textmining protocol.

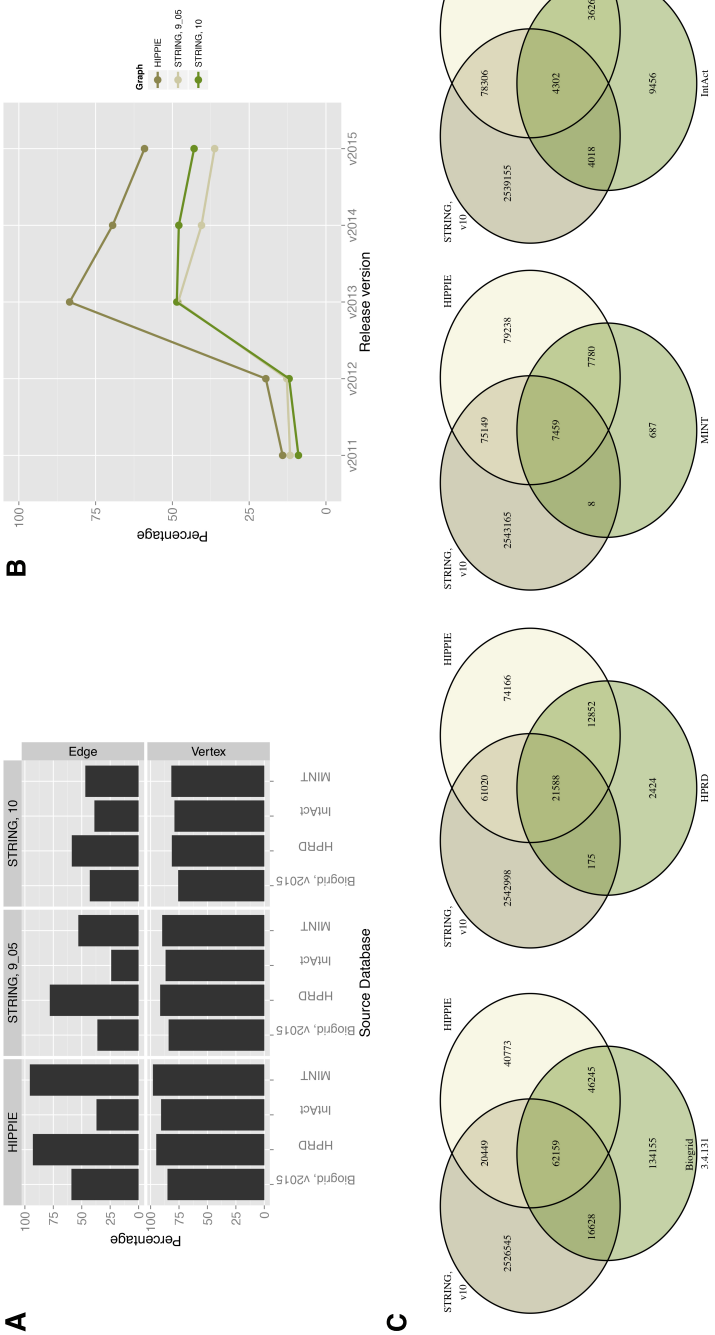


Figure 3: Edge inclusion patterns of STRING and HIPPIE from common primary database sources. Among edges from primary interaction databases, HPRD and MINT are almost fully included in HIPPIE, while only includes around 50% from these sources. Note that the inclusion patterns for the edges from primary interaction databases are not the same in v.9.05 and v.10 of STRING (A). Edge inclusion of BioGrid in HIPPIE and STRING as a function of BioGrid release year (B). The graph indicates that among BioGrid releases, the 2012 version has the most number of overlaps with all PPIs studied. Graph intersection checks (C) indicate that not only different proportions, but also different components of the primary interaction databases were incorporated in STRING and HIPPIE.

References

[1] Renate Konig, Chih-yuan Chiang, Buu P. Tu, S. Frank Yan, Paul D. DeJesus, Angelica Romero, Tobias Bergauer, Anthony Orth, Ute Krueger, Yingyao Zhou, and Sumit K. Chanda. A probability-based approach for the analysis of large-scale rnai screens. *Nat Meth*, 4(10):847–849, Oct 2007.

[2] Shashank Tripathi, Marie O Pohl, Yingyao Zhou, Ariel Rodriguez-Frandsen, Guojun Wang, David A Stein, Hong M Moulton, Paul DeJesus, Jianwei Che, LubbertusC F. Mulder, Emilio Yángüez, Dario Andenmatten, Lars Pache, Balaji Manicassamy, Randy A Albrecht, Maria G Gonzalez, Quy Nguyen, Abraham Brass, Stephen Elledge, Michael White, Sagi Shapira, Nir Hacohen, Alexander Karlas, Thomas F Meyer, Michael Shales, Andre Gatorano, Jeffrey R Johnson, Gwen Jang, Tasha Johnson, Erik Verschueren, Doug Sanders, Nevan Krogan, Megan Shaw, Renate König, Silke Stertz, Adolfo García-Sastre, and Sumit K Chanda. Meta- and orthogonal integration of influenza "omics" data defines a role for ubr4 in virus budding. *Cell Host & Microbe*, 18(6):723–735, November 2016.

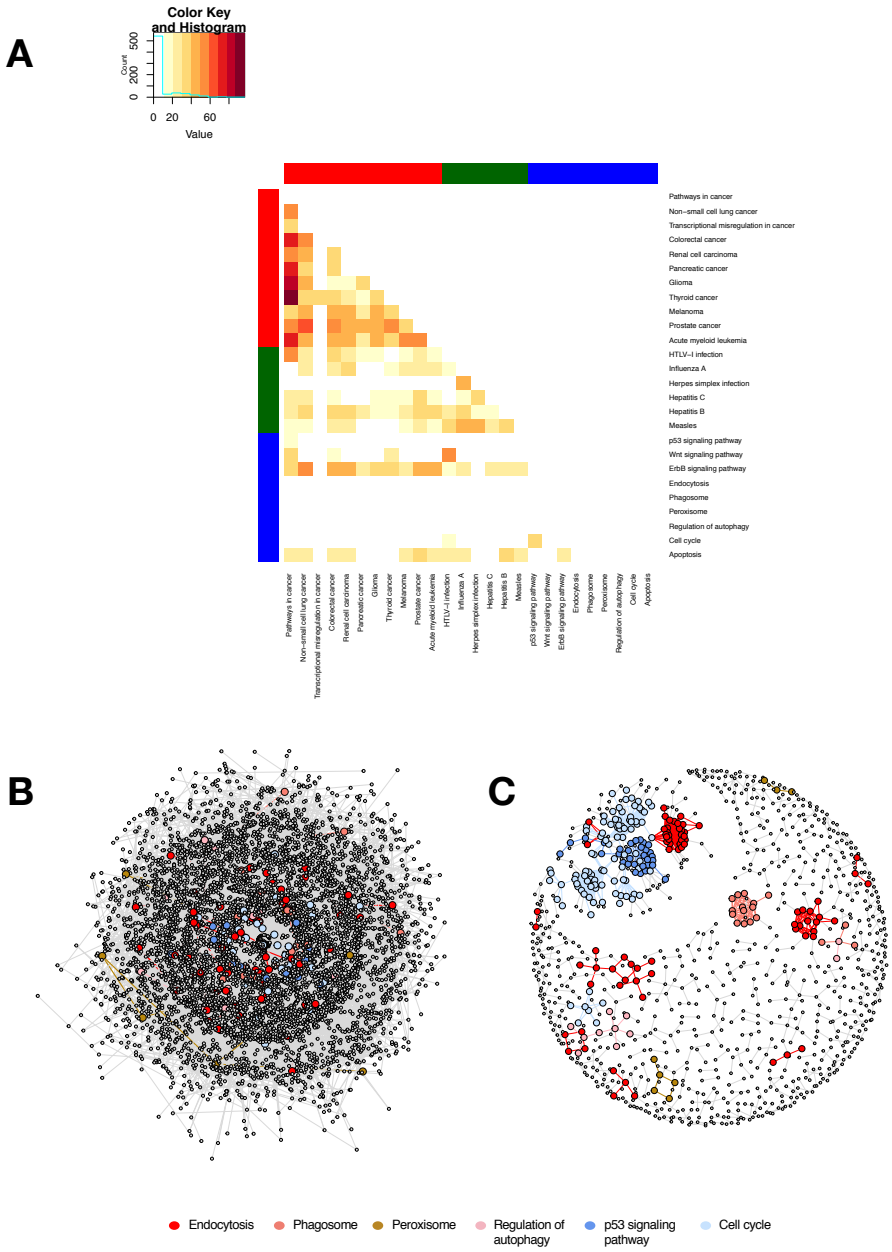


Figure 4: Proportion of overlapping edges in various KEGG networks (A). Six networks with minimal edge overlaps in (A) were selected and combined with its network neighborhood in the full STRING graph (B) for testing the combined score filtering. The corresponding confidence-filtered network (combined score > 800) shows the combined KEGG networks more distinctly, but would not be sufficient to separate the subgraphs; note that some of the subgraphs tend to form multiple, rather than concentrated clusters, and may not necessarily be retrievable using a combination of confidence filtering and community detection protocols.

Combined keywords in search = 2, min. match per edge = 1 Edges per pathway	POI	POI_CF	mPOI	Unretrieved uPOI	uPOI_e*	uPOI_w/oe*	uPOI_w/oe**	Extra ePOI	ePOI_e*	ePOI_w/oe	TP_recalc	FP_recalc	FN_recalc	recall	precision
Endocytosis	285	334	65	220	9	41	170	269	15	254	74	254	9	89.1566265	22.56098
Phagosome	72	293	9	63	0	22	41	284	7	277	9	277	0	100	3.146853
Peroxisome	7	50	3	4	4	0	0	47	8	39	7	39	4	63.6363636	15.21739
Regulation of autophag	9	37	6	3	3	0	0	31	8	23	9	23	3	75	28.125
p53 signaling pathway	60	410	53	7	1	0	6	357	105	252	54	252	1	98.1818182	17.64706
Cell cycle	360	469	195	165	50	27	88	274	65	209	245	209	50	83.0508475	53.96476

Vertices per pathway	POI	POI_CF	mPOI	Unretrieved uPOI	uPOI_e*	uPOI_w/oe*	uPOI_w/oe	Extra ePOI	ePOI_e*	ePOI_w/oe	TP	FP	FN	recall	precision
Endocytosis	84	421	48	36	22	14	NA	373	175	198	70	198	22	76.0869565	26.1194
Phagosome	21	322	6	15	3	12	NA	316	51	265	9	265	3	75	3.284672
Peroxisome	8	71	4	4	4	0	NA	67	54	13	8	13	4	66.6666667	38.09524
Regulation of autophag	11	54	9	2	2	0	NA	45	34	11	11	11	2	84.6153846	50
p53 signaling pathway	54	436	52	2	2	0	NA	384	280	104	54	104	2	96.4285714	34.17722
Cell cycle	95	413	82	13	13	0	NA	331	292	39	95	39	13	87.962963	70.89552

Combined keywords in search = 2, min. match per edge = 2 Edges per pathway	POI	POI_CF	mPOI	Unretrieved uPOI	uPOI_e*	uPOI_w/oe*	uPOI_w/oe**	Extra ePOI	ePOI_e*	ePOI_w/oe	TP	FP	FN	recall	precision
Endocytosis	285	117	35	250	12	41	197	82	9	73	47	73	12	54.6511628	31.97279
Phagosome	72	101	5	67	1	22	44	96	7	89	6	89	1	60	6.122449
Peroxisome	7	13	3	4	4	0	0	10	0	10	7	10	4	63.6363636	41.17647
Regulation of autophag	9	18	6	3	3	0	0	12	4	8	9	8	3	75	52.94118
p53 signaling pathway	60	206	50	10	2	0	8	156	66	90	52	90	2	92.8571429	36.11111
Cell cycle	360	247	167	193	63	27	103	80	27	53	230	53	63	74.6753247	77.18121

Vertices per pathway	POI	POI_CF	mPOI	Unretrieved uPOI	uPOI_e*	uPOI_w/oe*	uPOI_w/oe	Extra ePOI	ePOI_e*	ePOI_w/oe	TP	FP	FN	recall	precision
Endocytosis	84	157	39	45	14	31	NA	118	72	46	53	46	14	63.0952381	45.68966
Phagosome	21	149	8	13	1	12	NA	141	39	102	9	102	1	90	8.108108
Peroxisome	8	20	4	4	1	3	NA	16	10	6	5	6	1	55.5555556	35.71429
Regulation of autophag	11	25	9	2	1	1	NA	16	16	0	10	0	1	83.3333333	90.90909
p53 signaling pathway	54	202	49	5	3	2	NA	153	132	21	52	21	3	91.2280702	69.33333
Cell cycle	95	165	77	18	5	13	NA	88	82	6	82	6	5	82	81.18812

precision = TP/TP+FP
recall = TP/TP+FN
* with evidence from the orthogonal search
** with no evidence in STRING matching the search crit Includes cases when there are no matching literature results
NA: STRING evidence restricted to edges

Supplementary Table 5. Detailed context score filtering performance in selected KEGG pathways